

Bioinformatic approaches to regulatory genomics and epigenomics

376-1347-00L | week 09

Pierre-Luc Germain

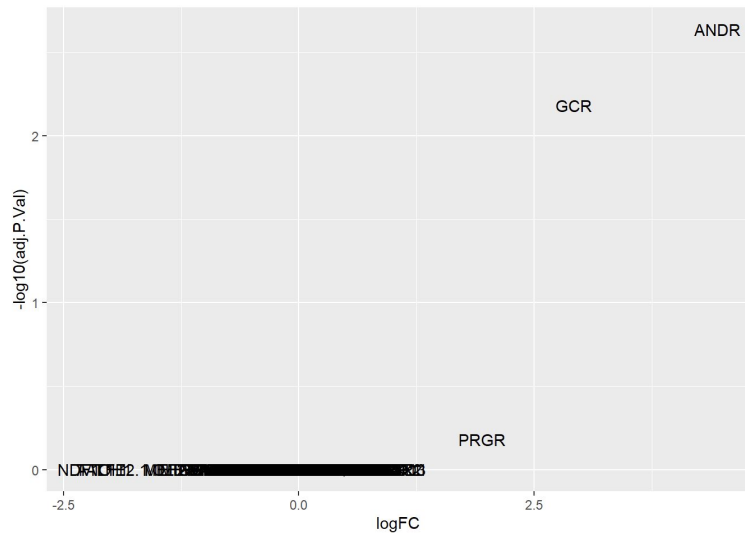
Plan for today

- Debriefing on the assignment
- Quick primer on count-based (discrete) statistics
- Our case study for the practical: GDVS
- Differential binding analysis
- Normalization methods

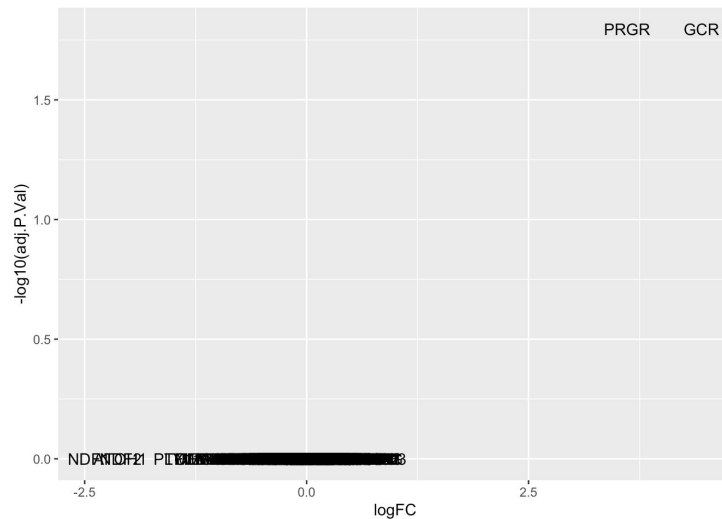
Debriefing on the assignments

Check species!

```
motifs <- query(MotifDb, c("Mmusculus", "HOCOMOCov10"))
```



```
motifs <- MotifDb::query(MotifDb, c("HOCOMOCov11-core", "Hsapiens"))  
length(motifs)
```



Debriefing on the assignments

Statistical testing:

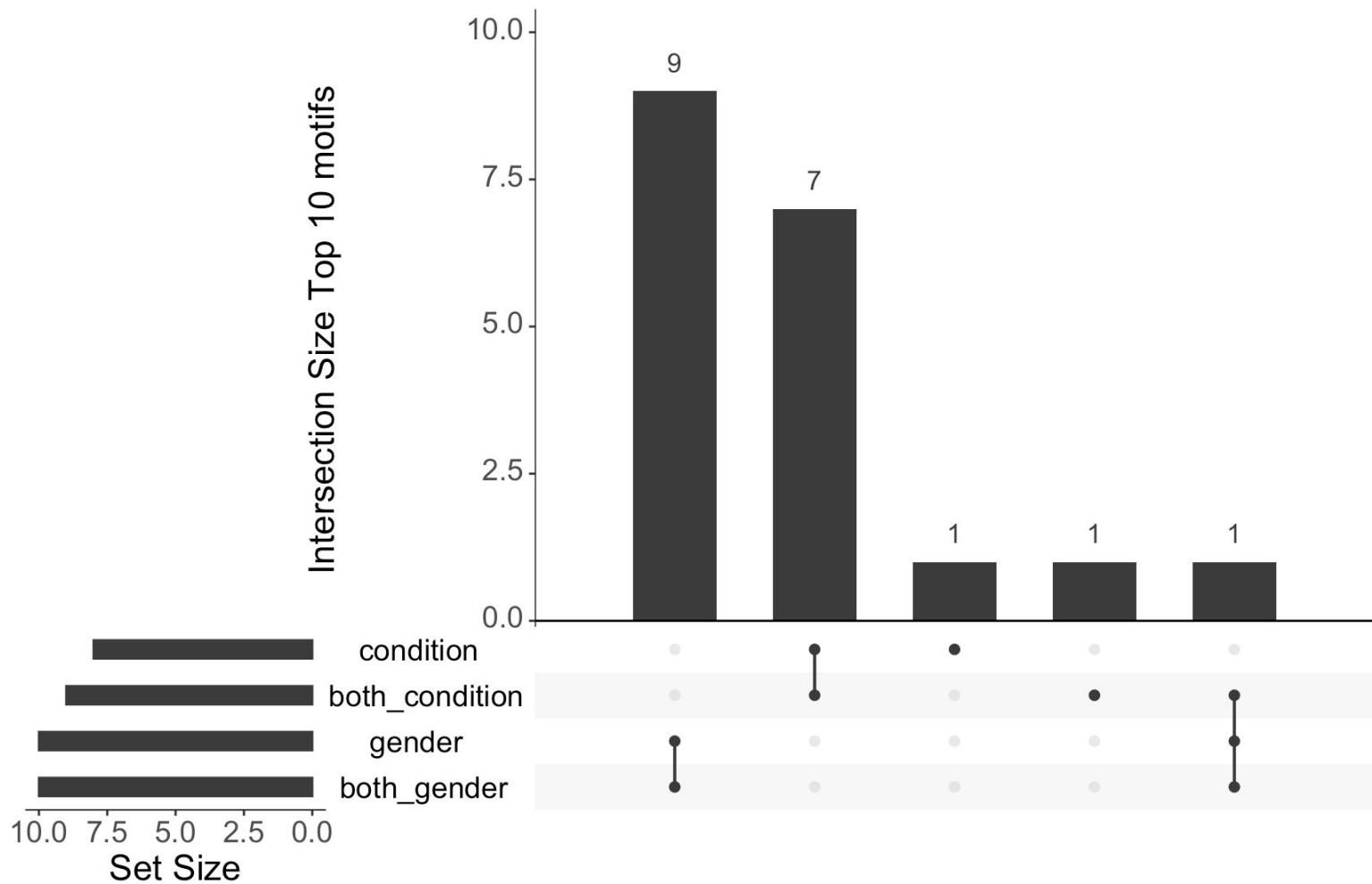
```
mm <- model.matrix(~sex+condition, ...)
```

and

```
mm <- model.matrix(~sex, ...)
```

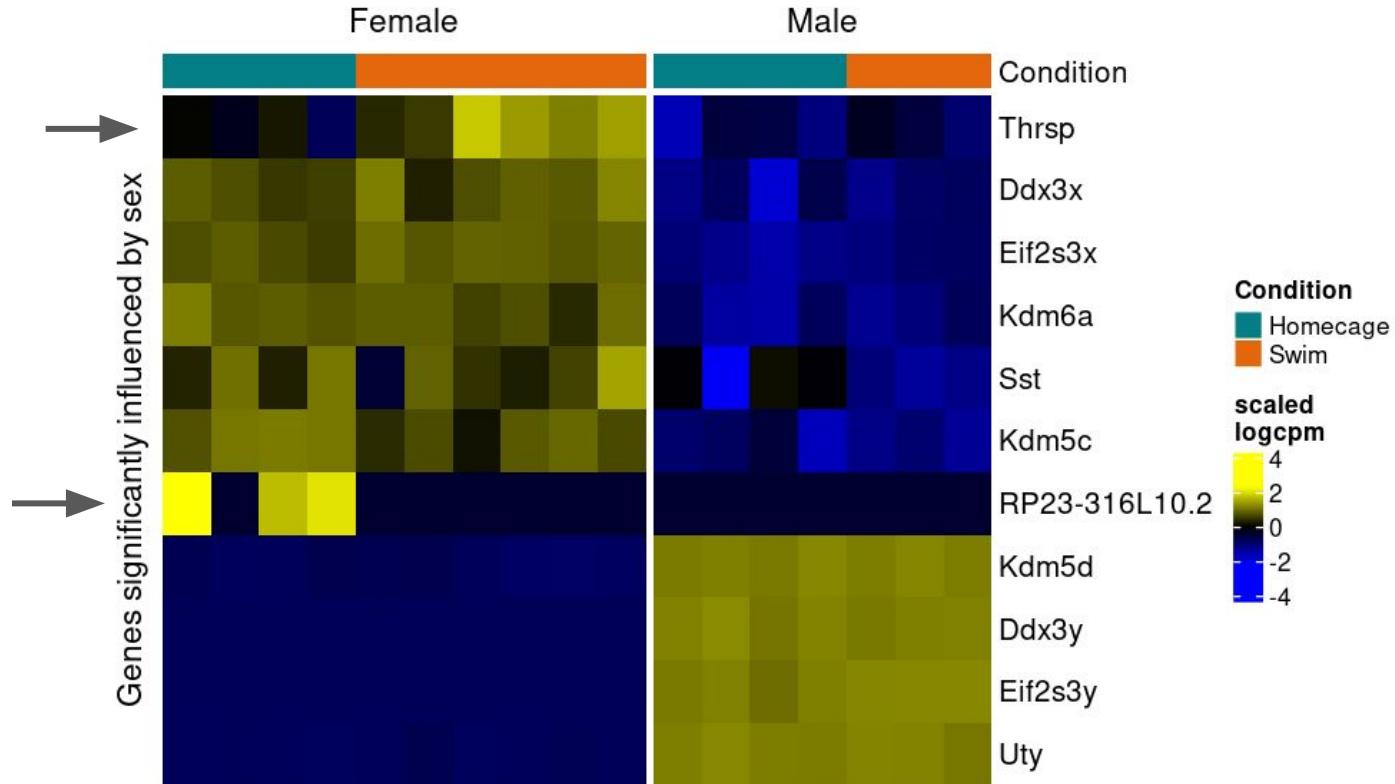
```
mm <- model.matrix(~condition, ...)
```

is not equivalent!



An example (from RNAseq)

Since some genes are stress-responsive, if the model does not include the condition the variability might appear high, leading to a failure to detect sex differences.



Count-based statistics

Count-based statistics

The **number of fragments** overlapping a given peak is not a **continuous** variable, but a **discrete** variable: it can only be non-negative integers.

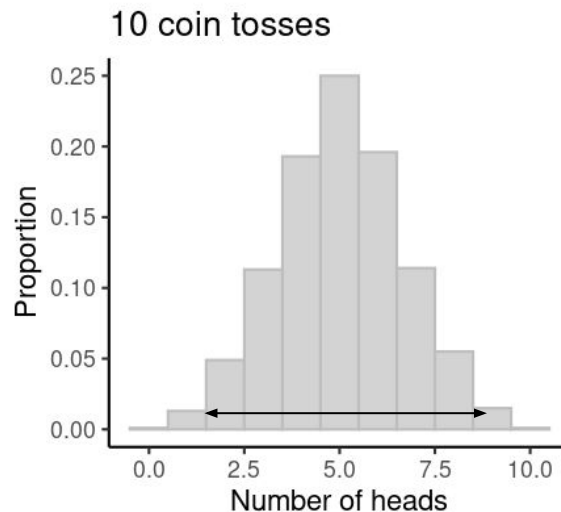
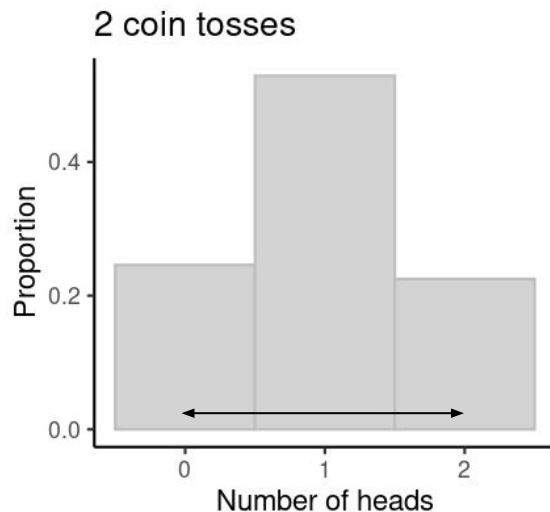
We must consider this information in the statistical analysis of count data, and so we can't use statistics based on a normal distribution (e.g. t-tests or standard linear models).

We'll use **generalized linear models**, which can be applied on different types of data distributions.

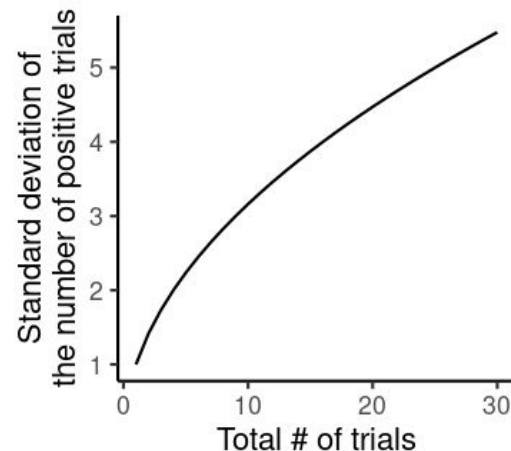
Count-based statistics

Consider coin tossing, where each toss can result in two outcomes.

Suppose each of us toss a coin X times, and we count the number of 'heads'.



The results are Poisson-distributed, meaning that the variance is equal to the number of tosses



Count-based statistics

However, counts coming from real phenomena (more complicated than coin tossing) are typically not quite Poisson-distributed: the variance typically increases *more* than would be expected, in other words they show **over-dispersion**.

This is typically because of sources of variation other than the random sampling process, which are typically unknown.

For this reason, count statistics are most often analyzed with the **negative binomial distribution**, which includes an extra dispersion parameter.

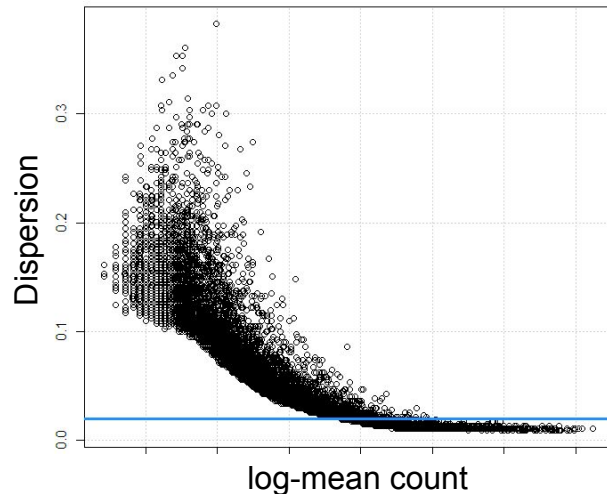
Information-sharing across features

In principle, each feature (e.g. gene or peak) we are analyzing could have a different degree over-dispersion, so that we would have to estimate this parameter separately for each of them.

Typically, we don't have enough samples to do this accurately.

RNAseq analysis packages such as [edgeR](#) therefore rely on a trend between average counts and their overdispersion to moderate those estimates.

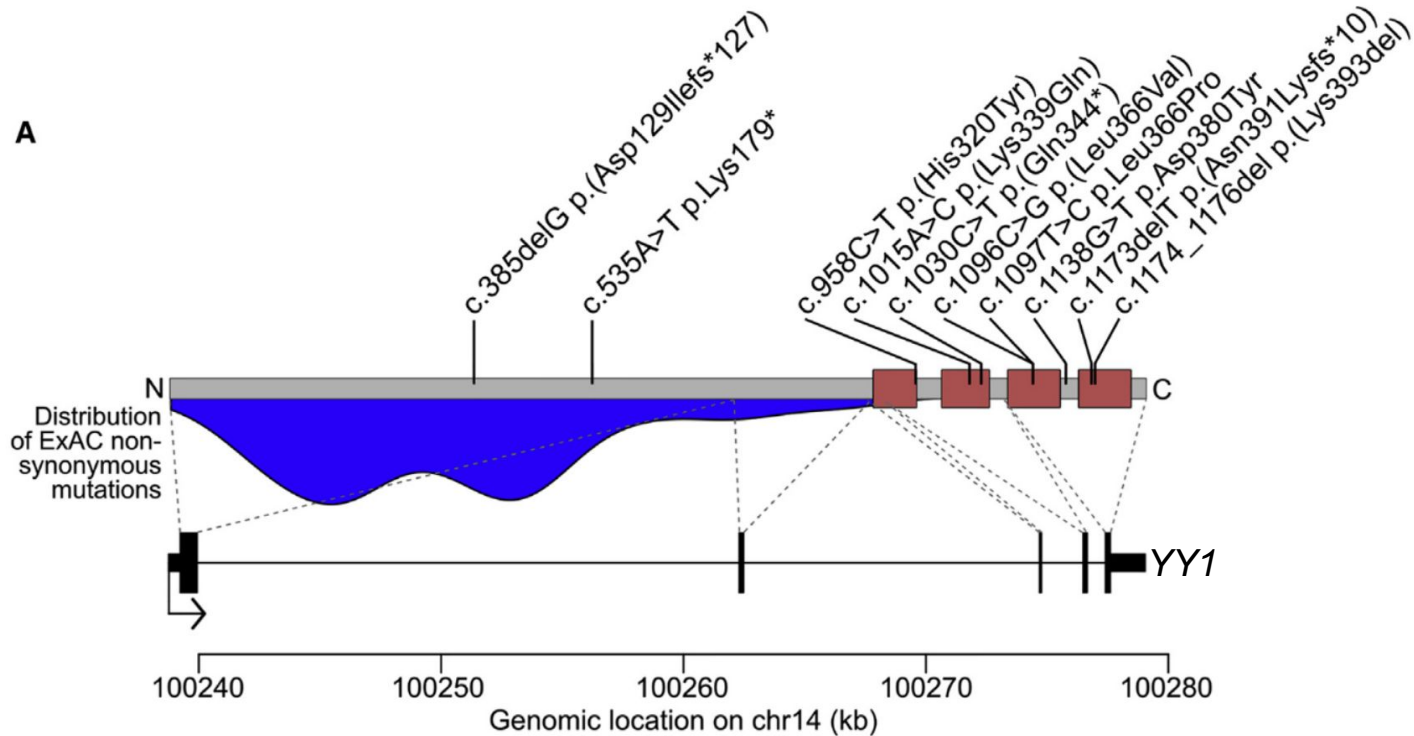
We can rely on the same packages for epigenomic data.



Our case study today

- Lymphoblastoid cells from patients with Gabriele-de Vries syndrome (and controls)
- OMIM:
 - “Gabriele-de Vries syndrome is an autosomal dominant neurodevelopmental disorder characterized by delayed psychomotor development, variable cognitive impairment, often with behavioral problems, feeding problems, some movement abnormalities, and dysmorphic facial features. Affected individuals may also have a variety of congenital abnormalities.”
- Caused by haploinsufficiency in the *YY1* gene
- Data: YY1 ChIP-seq in mutant and control lymphoblastoid lines

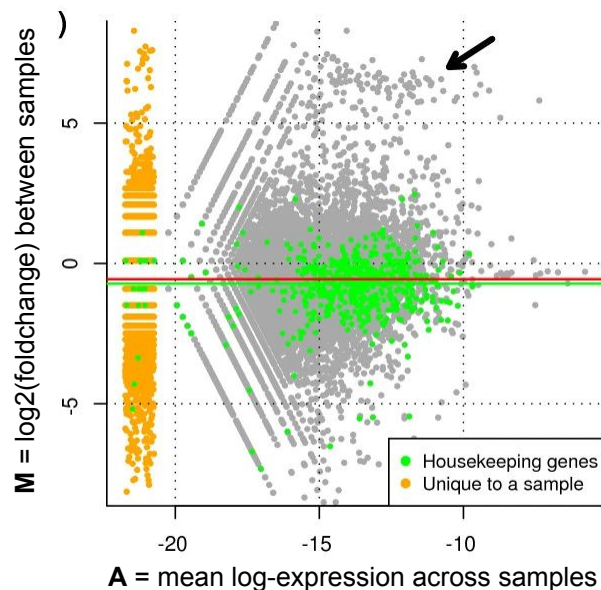
Our case study today



Practical

Normalization

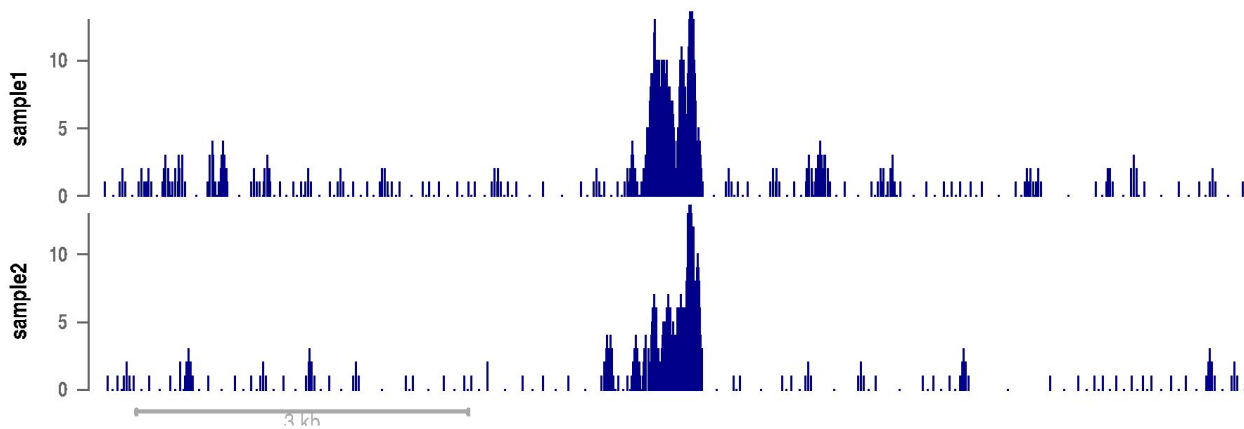
- Standard **TMM normalization** in edgeR assumes that the majority of regions don't change across condition.



(Adapted from Robinson & Oshlack, Genome Biology 2010)

Normalization

- Standard **TMM normalization** in edgeR assumes that the majority of regions don't change across condition.
 - This does not work when there are global increases or decreases (i.e. at most sites) between conditions
- **Background normalization** assumes that the background noise is the same across experiments

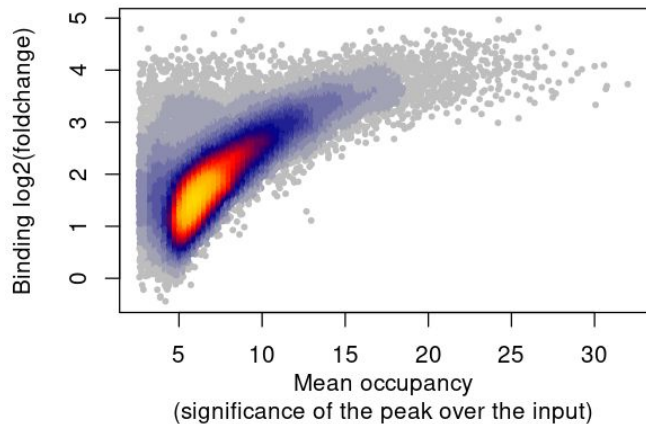


Normalization

- Standard **TMM normalization** in edgeR assumes that the majority of regions don't change across condition.
 - This does not work when there are global increases or decreases (i.e. at most sites) between conditions
- **Background normalization** assumes that the background noise is the same across experiments
 - This often works nicely, but can create artifactual differences in highly-enriched regions when the quality of the enrichment differs a lot between experiments
- Common (or top) peak normalization (e.g. **MAnorm**, Shao et al., Genome Biology 2012) assumes that the regions that are significantly enriched in both conditions don't change
- **S3norm** (Xiang et al., NAR 2020) performs both background and common peak normalization

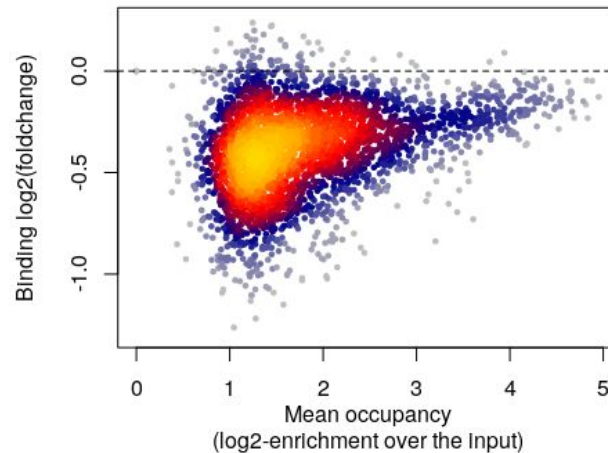
When the amount/activity of a TF is changed, not all sites change in binding equally

From inactive to active nuclear receptor



When activating a TF that wasn't active at all, the TF goes to high-occupancy sites, and these sites tend to show the largest changes in binding

YY1 haploinsufficiency



When reducing a TF that's already active, the TF also tends to keep going to the high-occupancy sites, so the low-occupancy sites are those that tend to show the largest change

* co-factors can influence these dynamics

Assignment

Until next week, come up with a preliminary plan for your project, summarizing:

1. What is the topic?
2. What data will you be using?
3. What are the analyses you wish to reproduce, or the questions you wish to answer?

This is not a final plan, but the start of a discussion!

Write that up in a Rmarkdown that you can render and upload to your repository.